GENERALISE NOT SPECIALISE: DESIGN IMPLICATIONS FOR A NATIONAL ASSESSMENT BANK

Rod Johnson and Sandra Johnson

Generalise not specialise: design implications for a national assessment bank

Rod Johnson and Sandra Johnson
Assessment Europe
mail@assessment.eu.com

Abstract

Within the framework of the Assessment is for Learning (AifL) programme[1], two systems of national assessment are currently operating in Scottish schools: on-demand 5-14 National Assessments and the sample-based Scottish Survey of Achievement. This paper will discuss issues surrounding the design of an assessment bank intended to support both systems.[2] It focuses in particular on the considerations underlying decisions about the structure of the shared materials database, the complex definition of an "item" that had to be adopted in order to accommodate a wide range of assessment types, the overall architecture of the wider information system, with its component databases (one being the bank) and information management subsystems, and the tensions arising from the need to accommodate the requirements of different systems of assessment while avoiding the dangers involved in data repetition and redundancy.

Introduction

Since the autumn of 2003, primary and lower secondary teachers in Scotland have benefited from online access to 'national assessments': these are tests which they can use on a voluntary basis to confirm their judgments about their pupils' levels of attainment in reading, writing and mathematics (for level descriptions see the relevant 5-14 curriculum guidelines: SOED 1991a for English language, 1991b and 1999 for mathematics)[3]. Schools make requests for assessments through a web interface[4], identifying their needs in terms of subject and level. A school might, for example, request a 'Level B' assessment in mathematics or a 'Level D' assessment in reading.

In reading, an assessment comprises two different tasks, where a task consists of a source text plus multiple associated test questions (20-30,

[1] www.ltscotland.org.uk/assess/
[2] There is, of course, an important third area of pupil assessment on a national scale in Scotland: external examinations, run by the Scottish Qualifications Authority (SQA). While the work discussed here is not mandated to cover application in the area of external examinations, we have tried as far as possible to keep the design we propose sufficiently flexible to accommodate this major category of high-stakes, pupil-based assessment.
[3] It is likely that at some point national assessments will be extended to include science and social subjects.
[4] www.aifl-na.net

depending on level). In mathematics, assessments comprise two loosely parallel 'booklets', each comprising 20-30 'atomistic' test items, all at the same level but spanning the mathematics curriculum at that level. Reading task pairs are selected at random from within a pool of appropriate assessment materials in response to individual requests. Mathematics booklets are created using domain sampling, i.e. random selections of items are drawn from within the materials store, following a test specification that dictates overall item numbers as well as imposing some constraints on content coverage.

The Scottish Survey of Achievement (SSA)[5], on the other hand, is a programme of annual sample-based surveys of pupil attainment at selected stages in primary and early secondary education. The SSA, launched in 2005, evolved from the Assessment of Achievement Programme (AAP), which was introduced in the mid-1980s and ran until 2004. The distinctive feature of the SSA is that pupil attainment is reported by individual local authorities as well as nationally, whereas the AAP reported only nationally. Attainment is currently reported for four subject areas, assessed on a 4-year rolling cycle – English language, mathematics, science and social subjects; core skills feature every year (reading, writing, numeracy, ICT, problem solving and working with others). In certain cases, domain sampling is employed to select items and to create tests for survey use. In all subjects, items are randomly allocated to pupils using multiple matrix sampling. Pupils' attainments are typically reported in terms of proportions attaining given 5-14 levels, using the same level descriptions as national assessments and the same decision criteria.

Both national assessments and the SSA assess pupils' attainments in essentially the same way, using the same kinds of assessment materials; indeed, materials used in the SSA are available post-survey for use in national assessments, and materials developed independently for use in national assessments are available also for survey use. Unsurprisingly, the decision was taken to maintain a shared resource of assessment materials, which we can call the 'assessment bank'. The assessment materials already in the bank[6], and others soon to be incorporated, are quite varied in nature, ranging from typical objective and short-answer forms to structured questions and themed item sets (e.g. reading tasks). Practical assessments of various types feature in the attainment surveys in all subject areas, and at some point these, too, will be banked.

But while the attainment surveys and national assessments draw largely on the same basic stock of assessment materials, the needs and aims of the two programmes are essentially different. One programme is pupil-based, and intended to provide teachers with information about their pupils to use when evaluating individual progress and determining next steps in learning; the other is cohort-based, where individual pupil assessment is subordinate to the gathering of information about the performance of the education system as a whole. These differences have significant implications for the structure and

[5] www.ltscotland.org.uk/assess/of/ssa/

[6] The Scottish Qualifications Authority is responsible for developing and maintaining bank content.

content of the assessment bank. In particular, it is important to maintain a perspective on the bank not as an isolated entity, but as one component of an evolving, larger and more complex information handling system targeted on assessment applications. We discuss the wider, dynamic context below, but first we need to consider the range of static information stored in the bank itself.

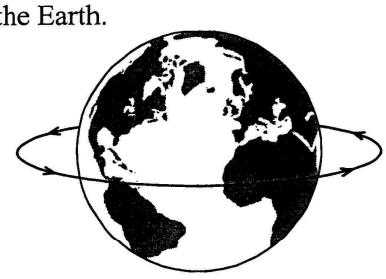
The assessment bank

Our design for the 5-14 national assessment bank, and for the SSA and national assessment information management systems, is based on several years' experience during the late 1990s/early 2000s, recovering historic AAP assessment materials and associated performance data and developing a prototype information management system for the programme (Johnson & Johnson, 2002 and 2003).

We came very early to the realisation that there was no simple organisational structure that would readily handle the wide variety of assessment materials used in the attainment surveys, in multiple subjects across a broad range of pupil ages. In particular, it was never going to be acceptable to design a banking system constrained to accept only objective format items[7], of which Figure 1 reproduces a typical example.

Source: AAP 2005a, Chapter 2, page 11

This is a picture of the Earth.



The Earth spins round and round. The arrows show it spinning.

How long does it take the Earth to spin round once?

Tick (✓) the correct answer.

1 hour ☐ 1 day ☐ 1 week ☐ 1 year ☐

Figure 1: A multiple-choice science item

[7] The developers of the national assessment bank in its present form failed to fully appreciate this, with the consequence that the bank now needs to be re-structured to accommodate the greater variety that was always present in the set of assessment materials used in past and current surveys.

Reading tasks offer the most extreme examples of assessment materials that do not fit the objective format item mould. These comprise a source text, or 'passage', followed by a relatively large set of questions, or items, grouped into 'sections', usually on the basis of a common format (see Figure 2).

Source: SSA 2006, Technical Annex, Section C, page C4

Stimulus text

In this example, *Attila the Hen*, a 420-word passage recalls events at Sunnycluck Farm just after all the hens have made their escape. Attila realises that the other hens are looking to her to lead them.

Section A: 10 multiple-choice questions

Section B

Arrange these sentences in the right order by putting the correct letters into the boxes below. The first one is done for you.*

A. The dogs hear Attila's squawk.

B. The hens return to the farmyard.

C. Attila leads the escape.*

D. The hens are upset by Attila's orders.

E. The men try to round up the hens.

F. A group of hens gather together.

*	1	2	3	4	5
C					

Section C

Here is a summary of part of the story after the farmyard battle. Fill each gap with **one or more words**. You may use words from the story or your own words.

Attila watched as the men returned to the _____ (1).

She decided to find out if _____ (2) had survived.

Taking a _____ (3) she _____ (4).

Soon she had assembled _____ (5).

Attila realised that her _____ (6), the old hen,

was _____ (7).

She decided she would have to _____ (8) the other hens

by herself because they _____ (9) her.

Figure 2: A typical reading task structure (abridged)

To handle this kind of assessment, we consider the typical unit of presentation in an assessment to be a *task*, perhaps with *subtasks*, containing *items*. An item is the smallest element of assessment with which we can associate a *score*.

It is, however, not always clear what exactly is the precise decomposition of a task into its constituent items. For example, does the task in Figure 3 contain a single item? Or three items? Or six?

Source: AAP 2005a, Chapter 2, page 10

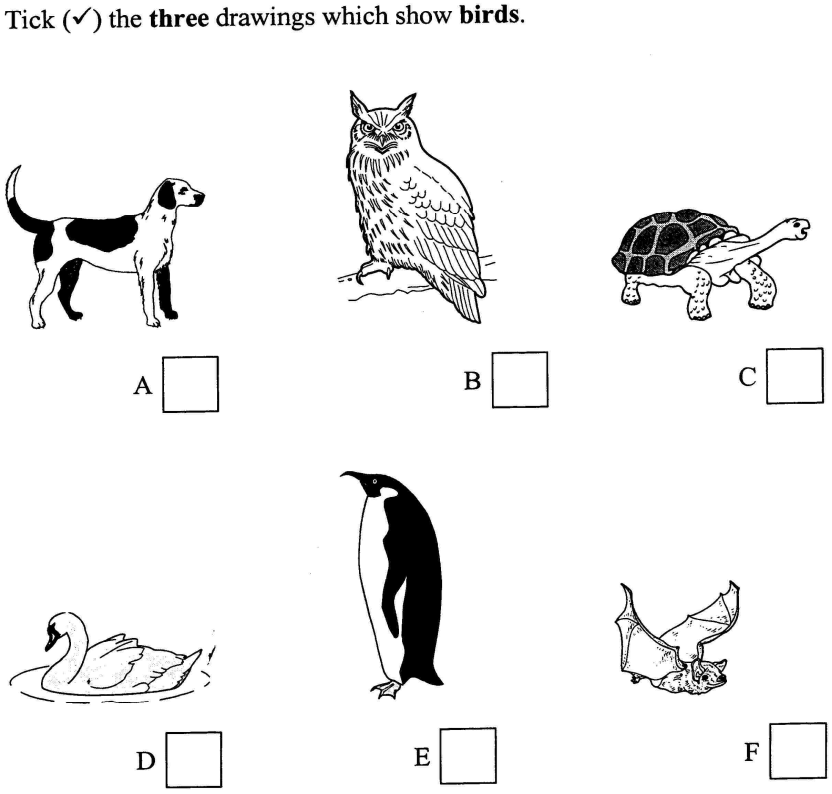


Figure 3: A science task comprising 'subitems'

Examples like this suggest that it may be useful to introduce a level of description below that of an item – a kind of atom to the item's molecule. Our design includes a notion of *subitem*, an element which can be associated with a pupil response, but which can only participate meaningfully in scoring its containing item when taken in conjunction with its fellow subitems.

Note that a response is not the same as a mark or score. The response is, ideally, the transcription of a subject's actual answer to the (sub)item, perhaps mapped to one or more of a finite set of possible responses, not all of which need to be correct; in the less ideal, but frequent, case where only information supplied by a marker is recorded, the response is just the marker-supplied information, again possibly mapped into a prescribed finite set. The (sub)item

mark is a binary quantity, representing the dichotomy correct/incorrect, derived by rule from the response: this is what we call the subitem *mark*. Where marker information only is recorded as the response, the relation between a response and its mark is just identity.

We define an item *score*, on the other hand, as a function of a set of subitem marks together with a rule for computing a composite numerical value from the responses, called a *mark scheme*. Subtasks and tasks also can have scores, usually computed by relatively trivial mark schemes (simple summation, for example).

The complete structure which we have currently implemented to handle the storage of the assessment materials is outlined in Table 1.

Task:	a set of questions, grouped into one or more *sections* or *subtasks*, normally based on a shared stimulus (text, picture, video clip …)
Subtask:	a collection of one or more *items*, based on the same stimulus material and usually, though not necessarily, sharing other common properties such as format, theme, level of difficulty; from the point of view of presentation subtasks are often labelled as *sections*; a subtask is characteristically the smallest unit of assessment whose external form can be independently stored
Item:	normally the smallest element of assessment which can be *scored*, though computation of the item score may involve consideration of *responses* to several constituent, usually interdependent, *subitems*
Subitem:	the smallest element of assessment for which a response can be recorded.

Table 1: A generalised ontology for storing assessment materials

In many types of assessment, an item and the corresponding task are expected to be equivalent (i.e. the task, subtask and item each contain just one component), the item has just one subitem, and all associated mark schemes are trivial, as is the case with orthodox multiple-choice items. Figure 1 above is an example, as is Figure 4 below, of what we often call a 'single-item' task.

Source: AAP 2005b, Chapter 2, page 9

Solve the following equation.

$$3(x-2)+7x=24$$

Answer: $x=$_____

Figure 4: A short answer mathematics item

This kind of item, however, in many assessment contexts, is very much a special case, and not the norm.

Consider the example in Figure 3 above: we treat this as a task having a single constituent item (and hence *a fortiori* just one subtask), the item having six subitems. The score for the item is a function of the set of responses to all six subitems (some of which may be blank).

In other examples, such as that shown in Figure 5, there is a clear composite structure, with a 'task' comprising (a single subtask with) two or more 'items'. Here the first item is a short answer question, while the second invites a more extended open-ended response. While the two items focus on the same general concept of force, they are in fact independent. Each could be presented quite separately, even without the introductory sentence and diagram (with a minor word change to the second item). But as they stand, from a presentational viewpoint there is little to be gained by storing them separately.

Source: AAP 2005a, Chapter 2, page 13

A spring balance can be used to measure the force exerted by something.

a. What is the unit of measurement of force?

_____

b. The box has a mass of 1kg but the scale on the instrument reads 10.

Why is this?

_____

_____

_____

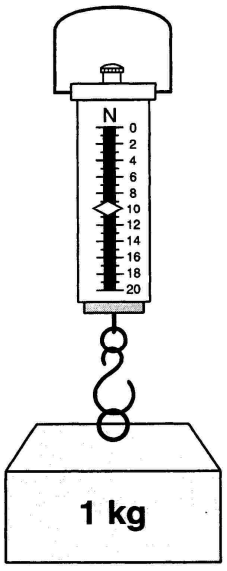


Figure 5: A composite 2-item science task

Figure 6 overviews a mathematical literacy task. This is a task typical of its kind, comprising a series of items based on a common stimulus. While independent in the sense that a correct answer to one item would not increase the chances of a correct answer to any other, the items in this case could not be presented separately from the others without reproducing all or the relevant part of the stimulus materials.

An interesting case of a multi-item reading subtask is that of a summary completion exercise (see Figure 2). Pupils are invited to fill gaps in a short summary of a longer text, implicitly reproducing the sense of the original whilst maintaining grammatical integrity. Here, each 'gap' is essentially a separate test item, but it would not be possible to present any item separately from the rest.

Source: AAP 2005b, Chapter 2, page 12

'Crime Survey'

The source material for this task comprises eight pie charts, illustrating the results of a survey into people's experience of crime. Each pie chart shows the proportion of individuals in the crime survey who answered in particular ways to questions such as "Have you, or another member of your immediate family, been a victim of crime in the last five years?" (response options: 'yes, self'; 'yes, other family member'; 'no'). Pupils are asked 12 questions, all requiring them to read information from one or other of the charts: five are short-response items, including "What percentage of people surveyed had **personally** had a crime committed against them in the last 5 years?" and seven are multiple-choice items.

Figure 6: Overview of a multi-item mathematical literacy task

Tasks, subtasks (to a lesser extent) and items have associated descriptive metadata, which we do not have space to go into here. Resource materials also have a set of associated descriptive metadata; where possible the resources themselves are incorporated into the bank.

A distributed information system

We said earlier that the assessment bank should be seen as just one component of a more complex architecture. To see why this would be so, recall that the materials in the bank should be directly available for use in at least two distinct contexts: the national assessments and the SSA.

While the system of national assessments is at present essentially a one-way communication system, in the medium term it is planned to develop the system further, in particular by facilitating 2-way communication, for example to receive pupil performance data from the schools, to provide feedback in the form of comparisons of class/school performance with national results (using SSA data), to allow pupils to take tests on-line, and/or to offer automatic marking to those teachers who request it.

For its part, administration of the SSA involves sampling from national school and pupil populations, communicating with authorities, schools and other

organisations, receiving, validating and processing pupil response data, carrying out automatic marking of item responses, producing a standard set of summative attainment reports, and keeping records of all of this activity as well as archiving response data at a detailed level for later retrieval for a variety of purposes.

It is evident that any attempt to incorporate one or the other of these functionalities directly into the bank design could risk prejudicing its utility for the other application. Moreover, the two functional descriptions above effectively describe the basic requirements for a pair of information management systems (IMS), respectively oriented towards the administration of on-demand test delivery and national system evaluation. These two observations together motivate our design for the union of the national assessments and the SSA into a distributed information system, based on a conceptual and organisational separation of the static, intrinsic characteristics of the materials themselves (the assessment bank proper) from dynamic, application-generated information (usage and performance data, *inter alia*, contained in dedicated information management subsystems).

A second shared resource, discussion of which is beyond the scope of this paper, is the set of externally maintained information about schools, information that is essential to the survey programme for sampling, distribution and analysis purposes and to the national assessments programme for the authentication and monitoring of requests from schools for assessments.

Figure 7 illustrates schematically the overall architecture of the system.

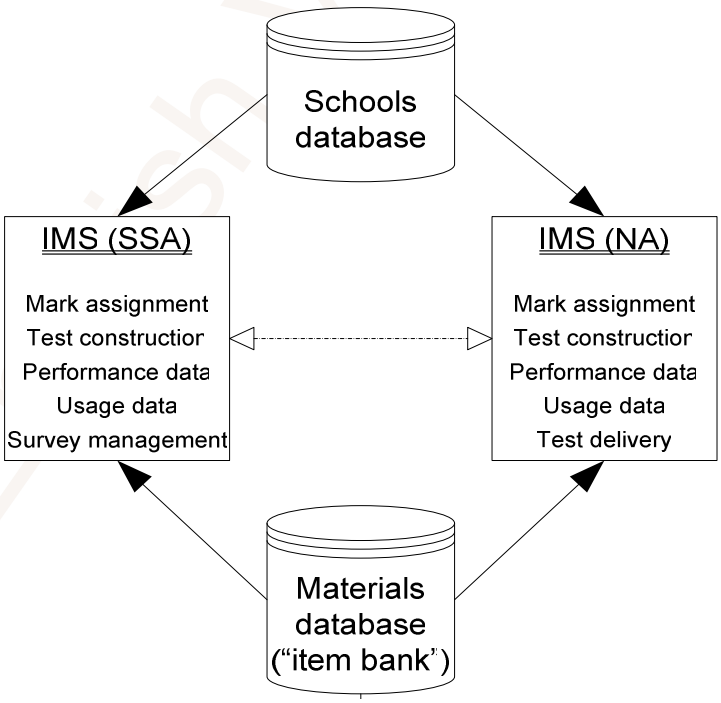


Figure 7: Distributed information system architecture

Note that information flows essentially in one direction from application-neutral databases to application-dependent IMS. At the same time, we would prefer

to minimise traffic between one IMS and the other, as symbolised in the diagram by the dotted line connecting the two, so as to allow as far as possible development to proceed independently.

As an example of the tension that can arise out of these constraints, consider the case where the developers of the national assessment IMS might choose to use item facility as part of a strategy for determining dynamically the balance of items in a test. Given that the same items are potentially used in SSA surveys, they would like to use relevant SSA performance data to produce the required facility estimates. So now the question arises as to where such data should reside, with the obvious temptation to store them directly alongside the items within the assessment bank. We are not in favour of this approach, for several reasons:

- such facility estimates are subject to dynamic change, as opposed to the stable, static information typically housed in the database;

- item facility is population-dependent, which means that facilities computed in one testing context might not be relevant for use in another;

- in any case, good data management systems design suggests that, *ceteris paribus*, values that can be readily computed from existing data should not be stored independently; if we follow this logic, we would have to consider storing the raw SSA responses themselves in the assessment bank; and if we store the SSA responses, why not the national assessments responses too?

- allowing the SSA IMS to deposit its results inside the assessment bank takes away control of the content of the bank from its own administrators.

On the basis of this kind of argumentation, we have had to conclude that a limited measure of interaction between constituent IMS has to be allowed, in order to maintain the autonomy and integrity of the assessment bank. Even so, we attempt to enforce the principle that all such interaction should always be subject to careful, bilateral negotiation.

Indeed, the question arises generally: what should form the content of the bank and what should more appropriately be located within the two dedicated IMS?

As we have just argued, we believe that item performance data appropriately belongs within the respective IMS, ideally in the raw form of pupils' qualitative responses to (sub)items, and not in the form of summative scores (these can at any time be generated on demand from the detailed response data). Similarly, usage statistics, those dynamic tracking statistics that monitor the use of individual items, tasks and tests, should also reside within each applications-specific IMS.

On the other hand, we have through experience come to the conclusion that, in addition to the assessment materials and associated descriptive metadata, the bank itself should hold a set of response options for each (sub)item, where

a set might comprise a single 'right answer', multiple alternative right answers or a series of right and wrong answers (which could have diagnostic value).

Mark allocations, however, and the construction of marking algorithms and mark schemes, should, in our view, more properly be held in some appropriate form within the applications-specific IMS. This is because mark allocations, even in the case of clearly identifiable individual items, can vary, depending on the purposes of the assessment application, the subject being assessed, the ages of the pupils/students being assessed, and the predilection of the assessors involved.

Test generation, too, should in our view reside within the IMS and not within the assessment bank. Again, this is because different applications can, and in our case do, use the same assessment materials to create quite different types of test or to create similar tests packaged differently.

In the national assessments, for example, mathematics tests are produced by drawing random samples of mathematics items from within the assessment bank, to provide an agreed representation of the curriculum, but with all the items at the same 5-14 level. In the recent 2005 SSA, numeracy tests were created in a similar way, but this time each test included items at three different 5-14 levels, with a randomised item ordering within the test itself. In both application areas the test generation algorithm might also change over time, another reason to keep this facility within each applications-specific IMS.

Finally, of course, each IMS is designed to deal with all the administration and transactions which characterise its particular application. In the case of the SSA, for example, the IMS should be expected to handle, *inter alia*, the generation of form letters to authorities and schools involved in the survey, specialised analyses of the results, and production of routine tables and reports.

In conclusion

The overall picture is one of some complexity, far greater than can be accommodated by the homogeneous, monolithic structure we might expect to find in conventional 'item banks', of the type implied in the oft-quoted definition (Sclater & McDonald 2004):

> "A collection of items for a particular assessment, subject or educational sector, classified by metadata which facilitates searching and automated test creation."

After several years of maintaining an archive of AAP survey materials and results, when faced with the challenge of designing an integrated resource which would serve adequately both the AAP's successor, the SSA, and a system of nationally available on-demand assessments, we concluded that the appropriate architecture was not an item bank as generally understood, but a distributed information system.

The system draws on the materials stored in an assessment bank as well as on shared information about schools and pupils. The bank is designed to

accommodate the wide variety of items and tasks that continue to be favoured by test developers in the different subject areas, to address assessment requirements at different levels in the education system, and to serve the specific needs of the different application domains. These are the two senses in which we have generalisation.

At the same time, we should be extremely careful not to bias the bank by imposing structures and behaviours which are largely the preserve of one application area, perhaps even to the extent of being in conflict with the needs of the other. Finally, we strive in our design, insofar as we are able, not to prejudice future extension of the bank to other, distinct applications (certification and selection, for example). In designing and developing such a complex artefact, tensions are bound to arise between, on the one hand, the conflicting requirements of different assessment needs within the system, and, on the other, the desire to maintain as high a degree as possible of application neutrality in the bank. Wherever possible we have tried to use principles of sound system design to resolve such tensions.

References

AAP (2005a). *The Sixth AAP Survey of Science (2003)*. Edinburgh: Scottish Executive Education Department.

AAP (2005b). *The Seventh AAP Survey of Mathematics (2004)*. Edinburgh: Scottish Executive Education Department.

Johnson, S. & R. (2002). *An architecture for the 5-14 Assessment Information System*. Internal report to the Scottish Executive Education Department, December 2002.

Johnson, S. & R. (2003). *A suggested basic structure for the National Assessment Bank*. Internal report to the Scottish Executive Education Department, April 2003.

Sclater, N. (2004), ed. *Item Banks Infrastructure Study (IBIS)*. www.toia.ac.uk/ibis.

Sclater, N., McDonald, M. (2004). 'Developing a national item bank', *Proceedings of the Eighth International Computer Assisted Assessment Conference*, Loughborough University, July 2004; cited in Niall Sclater (2004), ed.

SOED (1991a). *National Guidelines: English Language 5-14*. Edinburgh: Scottish Office Education Department.

SOED (1991b). *National Guidelines: Mathematics 5-14.* Edinburgh: Scottish Office Education Department.

SOEID (1999). *National Guidelines: Mathematics 5-14 Level F*. Edinburgh: Scottish Office Education and Industry Department.

SSA (2006). *The First SSA Survey of English Language and Core Skills. Technical Annex*. Edinburgh: Scottish Executive Education Department (in press).